12-1-2019

Evolution patterns of Peg3 and H19-ICR

Joomyeong Kim
Louisiana State University

Recommended Citation

Kim, J. (2019). Evolution patterns of Peg3 and H19-ICR. *Genomics, 111* (6), 1713-1719. https://doi.org/10.1016/j.ygeno.2018.11.025

Genomics 111 (2019) 1713–1719

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Evolution patterns of Peg3 and H19-ICR

Joomyeong Kim

Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA

ARTICLE INFO

Keywords:
YY1
CTCF
Tandem repeats
ICR
Genomic imprinting

ABSTRACT

Mammalian imprinted domains are regulated through small genomic regions termed Imprinting Control Regions (ICRs). In the current study, the evolution patterns of the ICRs of Peg3 and H19-imprinted domains were analyzed using the genomic sequences derived from a large number of mammals. The results indicated that multiple YY1 and CTCF binding sites are localized within the Peg3 and H19-ICR in all the mammals tested. The numbers of YY1 and CTCF binding sites are variable among individual species, yet positively correlate with the presence of tandem repeats within the Peg3 and H19-ICRs. Thus, multiple YY1 and CTCF binding sites within the respective ICRs may have been maintained through tandem repeats/duplications. The unit lengths of tandem repeats are also non-random and locus-specific, 140 and 400 bp for the Peg3 and H19-ICRs. Overall, both Peg3 and H19-ICRs may have co-evolved with two unique features, multiple transcription factor binding sites and tandem repeats.

1. Introduction

In mammalian genomes, a small subset of genes is expressed mainly from one allele due to genomic imprinting, by which one allele is usually repressed by DNA methylation and histone modifications [1,2]. Imprinted genes are clustered in specific regions of chromosomes, thus forming imprinted domains. The imprinting of several genes in a given domain is regulated by small genomic regions, termed Imprinting Control Regions (ICRs) [1,2]. ICRs acquire allele-specific DNA methylation during gametogenesis, which is maintained throughout the lifetime after fertilization [1,2]. Two types of ICRs exist for mammalian imprinted domains: 3 paternally methylated ICRs *versus* at least 20 maternally methylated ICRs [3]. The first type includes the ICRs of *H19/Igf2*, *Gtl2/Dlk1* and *Rasgrf1* imprinted domains. These ICRs are localized in intergenic regions, and acquire DNA methylation during spermatogenesis. The second type includes the ICRs of *Gnas*, *Peg3*, *Kcnq1* and *Igf2r* imprinted domains. These ICRs are localized close to the promoter regions of the associated genes, and acquire DNA methylation during oogenesis. The genomic sequences of both types of ICRs are CpG-rich, as seen in the promoter regions of mammalian genes. Besides carrying gametic DNA methylation marks, these ICRs are known to play critical roles for the corresponding domains in somatic cells, for instance, the well-known insulator function observed from the ICR of the *H19/Igf2* imprinted domain [4,5]. Thus, mutations on ICRs usually results in the disruption of ICR functions and also become molecular bases for the human diseases associated with genomic imprinting [1,2].

Several transcription factors are known to bind to these ICRs, including YY1 to the ICR of the *Peg3* domain and CTCF to the ICR of the *H19/Igf2* domain [4–7]. YY1 has been shown to function as a transcriptional enhancer for the *Peg3* domain [8,9], and also as a protector against DNA methylation for the Peg3-ICR [9]. On the other hand, CTCF functions as an insulator controlling the interaction between the promoters and enhancers within the *H19/Igf2* domain [4,5]. CTCF has also been shown to function as a protector against DNA methylation for the H19-ICR [10,11]. The YY1 and CTCF binding sites within both ICRs share the following unusual features [12]. First, the numbers of the binding sites in each ICR range from 4 to 10 binding sites, which is much greater than those observed from the typical promoter regions of the other mammalian genes. Second, the orientation of the binding sites is always in the same direction. Third, the relative positions of these binding sites within both ICRs are variable among individual mammals. Interestingly, the binding sites also tend to be closely associated with tandem repeats that are found within both ICRs, suggesting the co-evolution of these binding sites with tandem repeats in each mammalian species. Overall, these unusual features are known to be unique to ICRs, thus most likely associated with the functions of ICRs.

The biological functions or reasons for why the ICRs maintain multiple YY1 or CTCF binding sites during mammalian evolution are currently unknown. In the current study, therefore, I characterized the evolution patterns of Peg3 and H19-ICR with a main focus on multiple YY1 and CTCF binding sites and also on tandem repeats. I compared the genomic sequences of both ICRs derived from more than 30 mammals. The results indicated that multiple YY1 and CTCF binding sites might

E-mail address: jkim@lsu.edu.

https://doi.org/10.1016/j.ygeno.2018.11.025
Received 11 August 2018; Received in revised form 18 November 2018; Accepted 24 November 2018
Available online 29 November 2018

have been maintained through tandem duplication/repeats during mammalian evolution.

2. Results

2.1. Unique features of Peg3 and H19-ICRs

The genomic sequences of Peg3 and H19-ICRs display the following features. First, the 4-kb genomic region of the Peg3-ICR encompasses a 1.5-kb bidirectional promoter region for *Peg3* and *Usp29* and also the 2.5-kb 1st intron of *Peg3* (Fig. 1A). In both human and mouse, the Peg3-ICR is filled with multiple YY1 binding sites with an identical orientation but with different spacing. The Peg3-ICR is also associated with tandem repeats, but the length and relative position of these repeats are variable between human and mouse. Second, the mouse H19-ICR encompasses a 2-kb genomic region localized 2-kb upstream of the transcription start site of *H19* (Fig. 1B). The H19-ICR is also filled with multiple CTCF binding sites with an identical orientation. The length of the human H19-ICR is thought to be 4 kb in length, which is different from that of the mouse. The tandem repeat is detectable only from the human H19-ICR, but not from the mouse H19-ICR. In fact, the entire 4-kb region of the human H19-ICR is made of multiple units of a particular tandem repeat. Overall, both Peg3 and H19-ICRs display a set of conserved features between the two species, including multiple transcription factor binding sites and their frequent association with tandem repeats. However, detailed inspection also revealed that these seemingly conserved features still display lineage-specific variations, including the numbers and relative positions of YY1 and CTCF binding sites and the presence/absence of tandem repeats within the ICRs. This suggests that the genomic sequences of both ICRs may have been evolving independently in each lineage. Thus, I decided to characterize the evolution patterns of both Peg3 and H19-ICRs by comparing the genomic sequences derived from a large number of mammals. A series of screens were subsequently performed using the mouse genomic sequences of the Peg3-ICR (chr7:6,679,213-6,683,130 in mm9) and *H19* (chr7:149,761,434-149,764,046 in mm9) as probes with the Blat program (http://genome.ucsc.edu/cgi-bin/hgBlat). This series of analyses identified 33 Peg3-ICR and 38 H19-ICR from mammalian genomes (Files S1 and S2).

2.2. YY1 binding sites and tandem repeats of the Peg3-ICR

The identified sequences of the Peg3-ICRs were analyzed in the following manner. Potential YY1 binding sites were first identified from the genomic sequence of each species' Peg3-ICR using a custom-made Perl script (File S3). This screening was performed with two binding motifs: the first motif was CCAT.TT with a dot representing any base at that position and the second motif was GGCGCCATCTT. The second motif was selected based on its frequent association with the Peg3-ICR, as shown in both human and mouse Peg3-ICR (Fig. 1A). It also contains one CpG dinucleotide, which is known to be frequently methylated and subsequently deaminated, resulting in mutational decay, either TpG or CpA in the mammalian genomes [13,14]. Thus, the first motif, CCAT.TT, without this CpG site was designed to detect potential YY1 binding sites that had decayed during evolution. The results from this series of analyses are available (File S4). The main observations are summarized as follows. The YY1 binding sites with the first and second motifs were found from all the species tested (blue and red bars in Fig. 2). The direction of all the identified YY1 binding sites was also identical, which is similar to those seen in both human and mouse Peg3-ICR (Fig. 1A). The numbers of YY1 binding sites with the first motif identified from each species range from 5 to 21, whereas the numbers of YY1 binding sites with the second motif ranges from 1 to 9. In each tested species, the numbers of YY1 binding sites with the first motif, the degenerated form lacking CpG, were always much greater than those with the second motif, suggesting that YY1 binding sites within the Peg3-ICR have constantly decayed and been replenished throughout the mammalian evolution.

Potential tandem repeats were also identified and characterized using the bl2seq program (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=align2seq). Detailed information regarding the total length, unit length, and relative position of each tandem repeat are available (File S4). The main observations are summarized as follows. First, the Peg3-ICR of the majority of species contains tandem repeats except the following 6 species: tarsier, rabbit, shrew, alpaca, megabat, and rock hyrax (bottom panel in Fig. 2). It is relevant to note that these species without tandem repeats tend to have smaller numbers of YY1 binding sites than those with tandem repeats. Second, the total length of tandem repeats ranges from 1.0 to 1.6 kb, and, interestingly, the length of each repeating unit for a given tandem repeat is always about 140-bp in length for each individual mammal. The relative position of tandem repeats is usually located in the middle of the 1st intron of *Peg3* except for the primates, in which the tandem repeats tend to be localized more proximal to the 1st exon. Third, the sequences of tandem repeats are not conserved between different lineages, which was demonstrated by performing dot-plot analyses with a concatenated sequence that had been made of two species' sequences. As shown in Fig. 3A, two tandem repeats are visible with square regions that are filled with a set of diagonal lines. The first one on bottom left represents a tandem repeat from the human Peg3-ICR, whereas the second one on top right represents a tandem repeat from the mouse Peg3-ICR. On the other hand, two sections visualizing cross-species comparison, one on bottom right and the other on top left, did not show any significant set of diagonal lines, suggesting no cross-species sequence conservation of the identified tandem repeats. This is also the case for the Peg3-ICR with no obvious tandem repeats, such as rabbit and rock hyrax (Fig. 3B). Finally, mapping of the identified YY1 binding sites within each Peg3-ICR revealed that many YY1 binding sites with the degenerated motif tend to be clustered within the tandem repeat regions than the remaining regions of the Peg3-ICR, the first motif indicated with vertical blue lines on Fig. 3A. Thus, tandem repeats may be the genomic regions where YY1 binding sites have been undergoing dynamic turnover during evolution. On the other hand, the Peg3-ICR of the species without tandem repeats tend to have less numbers of YY1 binding sites with more scattered distribution patterns than those with tandem repeats (Fig. 3B). Overall, this series of analyses indicate that mammalian Peg3-ICRs have variable numbers of YY1 binding sites with fast turnover rates during evolution, and that tandem repeats/duplication may be responsible for the fluctuating numbers of YY1 binding sites among individual species.

2.3. CTCF binding sites and tandem repeats of the H19-ICR

The identified genomic sequences of the H19-ICR were similarly analyzed as described for the *Peg3* domain. Potential CTCF binding sites were first identified from the genomic sequence of each species H19-ICR with the following two motifs: C[CT][AG]C..GG.[GA]GC[*AG*]G and CCGC..GG.GGC[*AG*]G (Fig. 4). The second motif was selected based on its close association with the H19-ICR of both human and mouse. In contrast, the first motif was chosen as a degenerated motif of the second motif, which has variable bases at the CpG site. The outcome of this series of screening has been summarized as a graph (Fig. 4), and also the raw data are available (File S5). The main conclusions are as follows. CTCF binding sites were found from all the species tested, and the orientation of the identified CTCF binding sites was all identical as shown in the H19-ICR of human and mouse (Fig. 1B). The number of identified CTCF binding sites range from 1 to 10. The highest numbers were observed from the H19-ICR of dog and kangaroo rat with 10 binding sites, whereas the smallest numbers were from naked mole rat and guinea pig with 1 or 2 binding sites. On average 4 CTCF binding sites were found within the H19-ICR. Also, two independent searches with the first and second motifs derived the same number of CTCF

binding sites from the majority of species, suggesting that the CTCF binding sites localized within the H19-ICR may be evolutionarily stable without the process of CpG decay. This is quite different from the fast turnover rate observed from the YY1 binding sites localized within the Peg3-ICR (Fig. 2).

Potential tandem repeats were also analyzed with a similar approach as demonstrated for the Peg3-ICR. According to the results, the H19-ICR of the majority of species contain tandem repeats except the following 9 species: mouse, rat, squirrel, naked mole rat, guinea pig, pig, alpaca, ferret, and mega bat (bottom panel on Fig. 4). The detailed information regarding the total length and unit length of the identified tandem repeats has been summarized and is available (File S5). The main observations are as follows. The total length of the identified tandem repeats ranges from 2 to 7 kb with the shortest one observed from cat and the longest one observed from kangaroo rat. On average, the 4-kb long tandem repeat was found in the 2-kb upstream region of the *H19* locus. Interestingly, the unit length is mostly 400-bp long among the majority of the tested species. This is similar to the tandem repeats associated with the Peg3-ICR, showing a uniform 140-bp-long unit length among all the species (Fig. 3 and File S4). Sequence conservation of the tandem repeats was also tested with dotplot analyses as demonstrated for the Peg3-ICR (Fig. 5). The tandem repeats are clearly visible within the H19-ICR of human and dog, bottom left and top right, respectively. However, these two repeats do not show any cross-species sequence conservation based on no obvious matches in the bottom right and top left sections (Fig. 5A). On the other hand, tandem repeats are not detectable in the case of the mouse H19-ICR, but visible from the cat H19-ICR with a 2-kb square region, which is much less obvious than those from human and dog (top right on Fig. 5B). Individual CTCF sites were also mapped within each H19-ICR, and the results are presented as vertical lines on top panel in Fig. 5. In the species with tandem repeats, the CTCF binding sites are scattered throughout the entire length of tandem repeats (Fig. 5A). In the species without tandem repeats, on the other hand, the CTCF binding sites tend to be clustered within the 2-kb genomic region that is localized 2-kb immediate upstream of *H19* (Fig. 5B). This is quite opposite to the patterns observed from the Peg3-ICR, showing a more scattered pattern of YY1 binding sites from the species without tandem repeats (Fig. 3B). This might be an indication that the relative positions of individual CTCF binding sites within the H19-ICR might have been more functionally constrained than the positions of YY1 binding sites within the Peg3-ICR. Overall, this series of analyses indicated that mammalian H19-ICRs contain on average 4 CTCF binding sites, and also that the H19-ICR has been co-evolved with tandem repeats throughout mammalian evolution.

3. Discussion

In the current study, the evolution patterns of Peg3 and H19-ICR were analyzed with the genomic sequences derived from a large number of mammals. According to the results, multiple YY1 and CTCF-binding sites are found within the Peg3 and H19-ICR of all the mammals tested. The numbers of YY1 and CTCF binding sites are variable among individual species, yet positively correlate with the presence of tandem repeats within the ICRs of *Peg3* and *H19*. This suggests that the maintenance of multiple YY1 and CTCF-binding sites within both ICRs may have been mediated through tandem repeats/duplications. Overall, both Peg3 and H19-ICRs may have co-evolved with two unique features, multiple transcription factor binding sites and tandem repeats.

The ICRs of *Peg3* and *H19* are closely associated with two features,

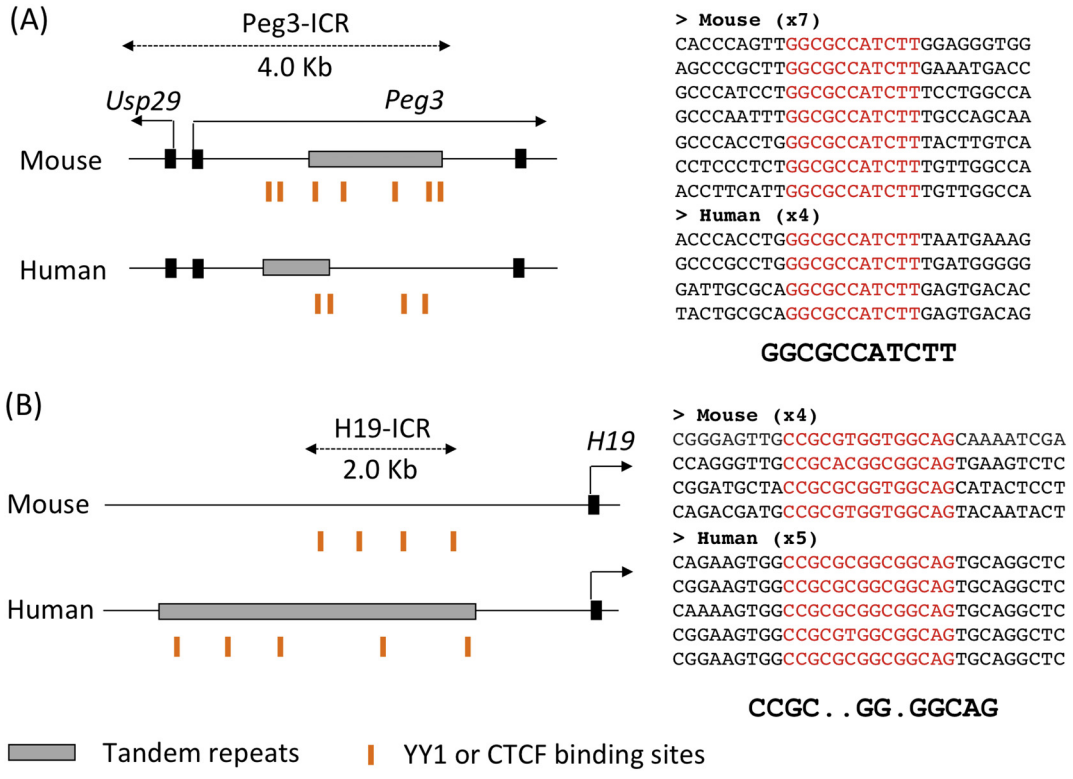

Fig. 1. Genomic structures of Peg3 and H19-ICRs. (A) Schematic representation of the 4-kb genomic interval of mouse and human Peg3-ICR. Transcriptional directions are indicated with arrows, while exons are indicated with filled boxes. The regions with tandem repeats are indicated with gray boxes. The positions of YY1 binding sites are indicated with vertical lines with orange. The sequences of YY1 binding sites are shown on right. The consensus sequence of YY1 sites is shown on bottom. (B) Schematic representation of the 7-kb upstream regions of mouse and human *H19*. The 2-kb region of mouse H19-ICR coincides with the genomic interval containing 4 CTCF binding sites, which are indicated with vertical lines with orange. The 4-kb genomic region of human H19-ICR is indicated with a gray box, since this region contains tandem repeats. The sequences of CTCF sites found within the mouse and human H19-ICRs are shown on right. The consensus sequence of CTCF sites is shown on bottom with a dot indicating any base at the position.

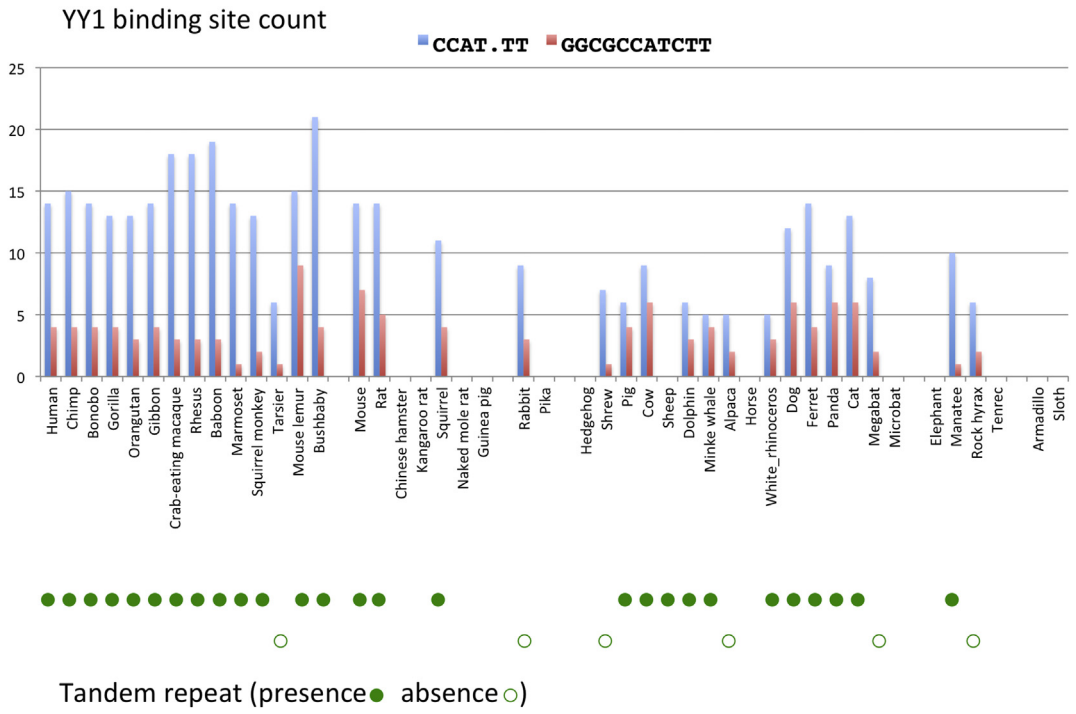

Fig. 2. Numbers of YY1 binding sites within the Peg3-ICR. The graph summarizes the number of YY1 binding sites identified from the Peg3-ICR of each mammalian species. The numbers of YY1 binding sites that are identical to the consensus sequence (GGCGCCATCTT) are represented with red bars, whereas the numbers of the YY1 binding sites matched with the degenerated sequences (CCAT.TT) with blue bars. The presence and absence of tandem repeats within the Peg3-ICR are indicated with filled and unfilled green circles, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

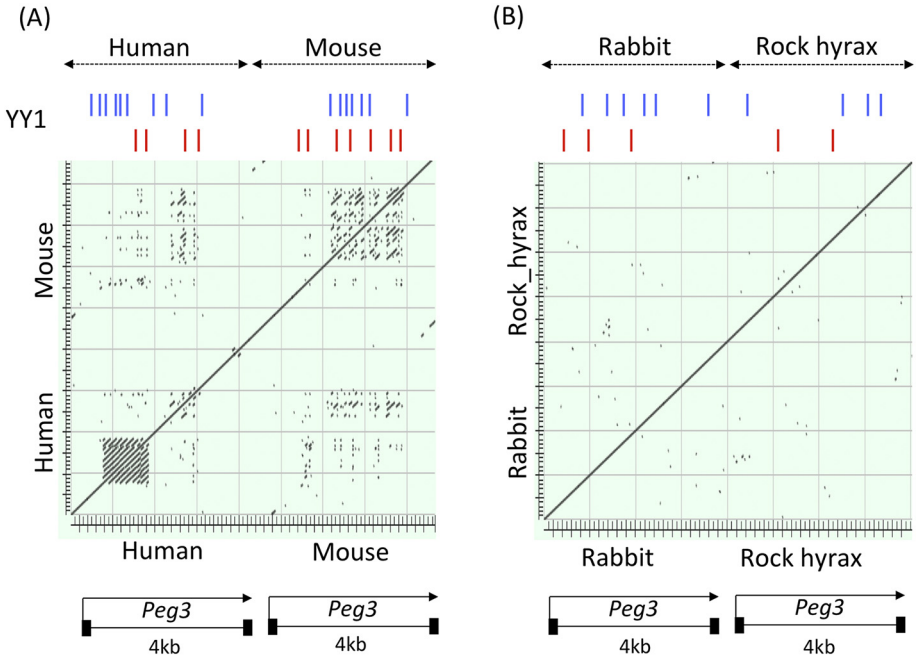

Fig. 3. Tandem repeats within the Peg3-ICR. (A) The 4-kb genomic sequences of human and mouse Peg3-ICR were first concatenated, and subsequently used for dotplot analyses. The subsequent results visualize genomic regions with tandem repeats, which are filled with a set of diagonal lines. The 1-kb region on bottom left and the 1.6-kb region on top right are identified as tandem repeat regions in the human and mouse Peg3-ICR, respectively. The results also visualize sequence conservation between two species, which are shown on bottom right and top left. These two sections do not show any square with clear diagonal lines, indicating that the tandem repeats detected from each species are species-specific without any sequence conservation. The positions of YY1 binding sites are indicated with vertical lines on top with orange and blue indicating the consensus and degenerated YY1 binding sites. The degenerated YY1 sites tend to be clustered within the tandem repeat regions, indicating the rapid turnover rates of YY1 binding sites. (B) The 4-kb genomic sequences of rabbit and rock hyrax Peg3-ICR were also concatermized and analyzed with dotplot analyses. Both species do not show any square with a set of diagonal lines, indicating no detectable tandem repeats within the tested genomic regions. Nevertheless, both regions are filled with YY1 binding sites, which are indicated with vertical lines with orange and blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

multiple transcription factor binding sites and tandem repeats (Fig. 1–5). These features are unique to these two ICRs, since similar features have not been observed from the other known ICRs. It is relevant to point out that the ICR of the *Gnas* domain was previously reported to have multiple YY1 binding sites, but the evolution pattern of this ICR was different from the ICRs of *Peg3* and *H19*: the numbers of YY1 binding sites are sporadic between lineages and, furthermore, no obvious tandem repeats in each lineage (File S6). In terms of evolution pattern, the two features detected within the ICRs of *Peg3* and *H19* also show lineage-specific variations. For instance, the numbers and relative positions of YY1 and CTCF binding sites are variable among individual species. Also, tandem repeats are detected from the majority, but not

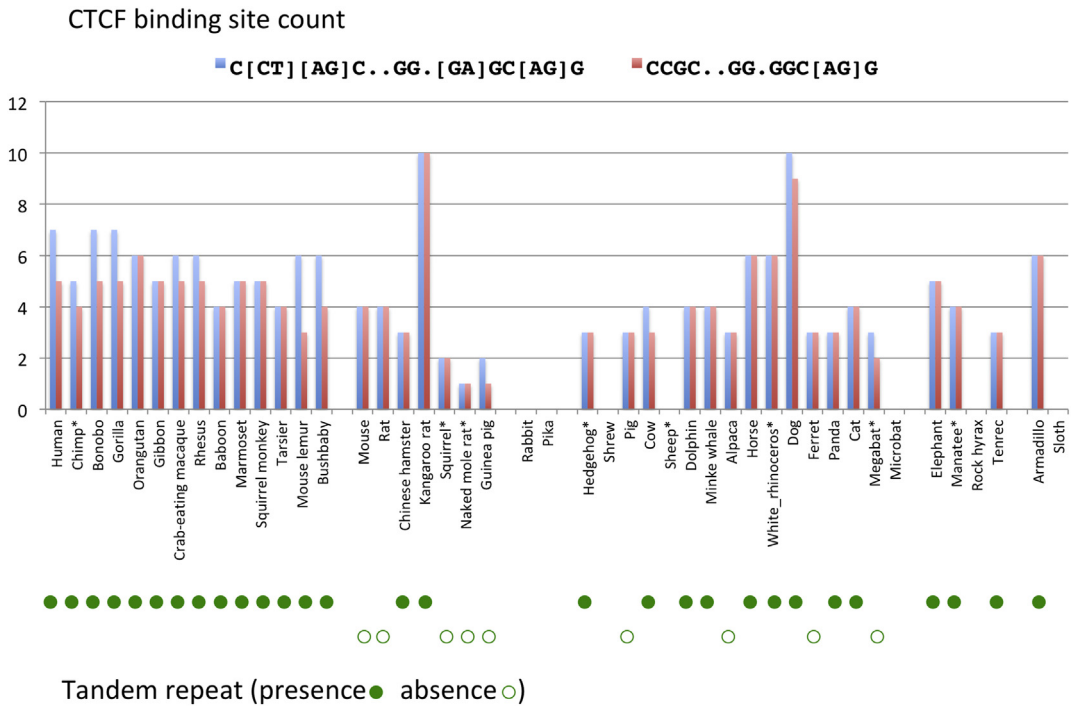

Fig. 4. Numbers of CTCF binding sites within the H19-ICR. The graph summarizes the number of CTCF binding sites identified from the H19-ICR of each mammalian species. The numbers of CTCF binding sites that are identical to the consensus sequence (CCGC..GG.GGC[*AG*]G) are presented with red bars, whereas the numbers of the CTCF binding sites matched with the degenerated sequences (C[CT][AG]C..GG.[GA]GC[*AG*]G) with blue bars. The presence and absence of tandem repeats within the H19-ICR are indicated with filled and unfilled green circles, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

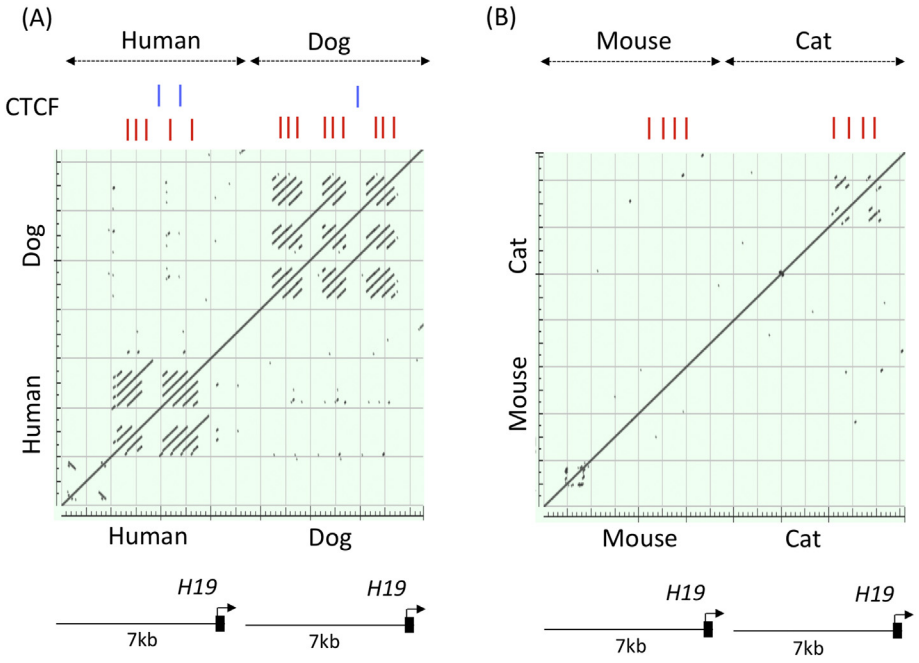

Fig. 5. Tandem repeats within the H19-ICR. (A) The 7-kb genomic sequences of human and dog H19-ICR were first concatenated, and subsequently used for dotplot analyses. The subsequent results visualize genomic regions with tandem repeats, which are filled with a set of diagonal lines. The 4-kb region on bottom left and the 5-kb region on top right are identified as tandem repeat regions in the human and dog H19-ICR, respectively. The results also visualize sequence conservation between two species, which are shown on bottom right and top left. These two sections do not show any square with clear diagonal lines, indicating that the tandem repeats detected from each species are lineage-specific without any sequence conservation. The positions of CTCF binding sites are indicated with vertical lines on top with orange and blue indicating the consensus and degenerated CTCF sites. (B) The 7-kb genomic sequences of mouse and cat H19-ICR were also concatermized and analyzed with dotplot analyses. The mouse region does not show any square with a set of diagonal lines, indicating no detectable tandem repeats within the tested genomic region. The cat region shows a 2-kb region with a much smaller number of diagonal lines than those detected from the human and dog regions, thus the tandem repeats identified from the cat region were much less obvious than the human and dog regions. The 2-kb upstream region of both mouse and cat H19-ICR are filled with CTCF sites, which are indicated with vertical lines with orange. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

all, of the mammals tested. Furthermore, the identified tandem repeats may have different evolutionary ages based on their different levels of sequence identity between individual species, suggesting that tandem duplications generating these tandem repeats may have occurred independently at different evolutionary times in individual lineages. The biological reasons for the frequent tandem duplications observed within the Peg3 and H19-ICRs are currently unknown, but might be related to multiple transcription factor binding sites. Several series of studies have already demonstrated that both YY1 and CTCF play pivotal roles for the ICR function of the *Peg3* and *H19*-imprinted domains: YY1 as a transcriptional enhancer for the *Peg3* domain and CTCF as an insulator for the *H19* domain [4,5,8,9]. Thus, these transcription factor-binding sites

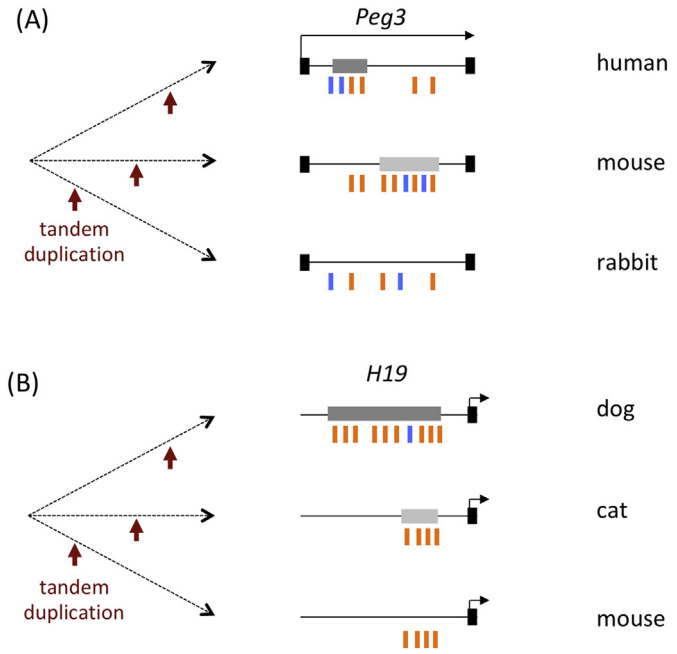

Fig. 6. Evolution patterns of Peg3 and H19-ICRs. Schematic representation of the evolution patterns associated with the Peg3 and H19-ICR (A, B). Tandem repeats are found within both ICRs, which are indicated with gray boxes. Tandem duplication events, indicated with vertical red arrows, may have occurred in a lineage-specific fashion at different evolution time points, resulting in the formation of species-specific tandem repeats with different evolutionary ages. The darker gray ones indicate the repeats with younger ages that have been duplicated more recent evolutionary times than the bright gray ones with older ages. The ICRs without any gray boxes may have lost the tandem repeats due to their older ages. The YY1 and CTCF binding sites are indicated with vertical lines with orange and blue colors, which indicate the perfectly matched and degenerated binding sites, respectively. In the case of the Peg3-ICR, many degenerated YY1 binding sites are found within tandem repeat regions, thus suggesting their rapid turnover during evolution. Given the distribution pattern of CTCF sites, it is most likely that tandem duplication has been involved in the formation of the entire length of the H19-ICR during mammalian evolution. In contrast, tandem duplication is believed to have been involved in duplicating only part of the Peg3-ICR. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

need to be maintained as key *cis*-regulatory elements within the ICRs of both domains. Yet, both transcription factor-binding sites are thought to be very vulnerable to degeneration, since both binding sites contain a CpG site, which is known to decay very fast in the mammalian genomes [13,14]. As a means of replenishing the degenerated sites, tandem duplication might have been adopted for the genomic regions of both ICRs. These tandem duplications are also predicted to have occurred at different evolutionary times in individual lineages. As a consequence, some lineages may have tandem duplication at the early stage of mammalian evolution, and gradually lost the sequence structure of tandem repeats (Peg3-ICR in rabbit and H19-ICR in mouse, Fig. 6), whereas the other lineages may have tandem duplications in relatively recent times, thus providing easily visible tandem repeats (Peg3-ICR in human and H19-ICR in dog, Fig. 6). These independent tandem duplications might also be responsible for the observed variations in the numbers and relative positions of YY1 and CTCF binding sites between different lineages. Overall, these two unique features, YY1 and CTCF binding sites and tandem repeats, may have co-evolved and contributed to the formation and maintenance of both Peg3 and H19-ICRs.

According to the results, there is one interesting difference between the two types of tandem repeats observed from the Peg3 and H19-ICRs. The unit length of the tandem repeats identified from the Peg3-ICR is about 140-bp long, whereas the unit length from the H19-ICR is about 400-bp long. These locus-specific unit lengths are particularly intriguing since the majority of tandem repeats in each locus have been generated multiple times at different evolutionary times. This implies that each locus may have allowed certain tandem duplications based on their unit lengths, although tandem duplication is expected to occur without any limitation on unit length. This may be related to the functions of each ICR. The Peg3-ICR encompasses a bidirectional promoter for *Peg3* and *Usp29*, thus is actively involved in transcription. Yet, the unit length of the tandem repeats identified from the Peg3-ICR appears to be 140 bp in length, which coincides roughly with the DNA length of one nucleosome unit [15]. Thus, this unit length might be closely associated with the nucleosome structure. Duplicating genomic regions on a 140-bp interval might be allowable without disrupting the already established nucleosome positioning of this promoter region. On the other hand, the H19-ICR is localized in the intergenic region, and functions as an insulator, which might require higher-order nucleosome/chromatin structure different from the Peg3-ICR. Thus, the larger unit length of the tandem repeats might be allowable without disrupting some unknown aspects of the insulator function of the H19-ICR. Overall, although speculative, it should be interesting to test these predictions in the near future.

Multiple YY1 and CTCF binding sites appear to be evolutionarily selected features within Peg3 and H19-ICRs, although the biological reasons are currently unknown. It is also unknown whether all the binding sites in a given species' ICR are functional, which warrants future experiments testing *in vivo* binding, such as ChIP (Chromatin Immunoprecipitation)-based surveys. Nevertheless, two possible scenarios have been previously proposed [12]. One possibility might be that the ICRs might need high levels of the YY1 or CTCF protein amounts. Multiple transcription factor binding sites might ensure attracting YY1 and CTCF proteins nearby the genomic regions of both ICRs. The other possibility might be that multiple transcription factor binding sites might be needed due to the relatively large genomic sizes of both ICRs as regulatory regions, ranging from 2 to 4 kb in length. Both possibilities are not mutually exclusive, thus could be together the reasons for maintaining multiple transcription factor binding sites for both ICRs. Yet, both binding sites are vulnerable to decay due to the frequent methylation on their CpG sites, thus tandem duplications might have been adopted as means of replenishing a sufficient number of transcription factor binding sites during mammalian evolution.

4. Materials and methods

4.1. Screening of Peg3 and H19-ICRs from mammalian genomes

This genomic sequences of mouse Peg3-ICR (chr7:6,679,213-6,683,130 in mm9) and *H19* (chr7:149,761,434-149,764,046 in mm9) were used as probes to identify the corresponding orthologous sequences from the other mammals. Screening was performed using the BLAT program (http://genome.ucsc.edu/cgi-bin/hgBlat). In the case of H19-ICR, the identified *H19* region was used as a position marker to identify its 7-kb upstream region as a potential region containing the corresponding H19-ICR. The identified sequences were formatted as fasta files, and used for the following analyses described below (Files S1 and S2).

4.2. YY1 and CTCF binding site detection and dotplot analysis

The identification of YY1 and CTCF binding sites was performed with a custom-made Perl script (File S3), and the results from this series of bioinformatics analyses are available (Files S4 and S5). The genomic sequence of a given species was analyzed to identify tandem repeats with the bl2seq program (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=align2seq). The subsequent results were inspected through dotplot analyses. The results have also been detailed along with the numbers of

YY1 and CTCF binding sites (Files S4 and S5).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.11.025.

Acknowledgements

I would like to thank Dr. Bambarendage Pini Perera and Subash Ghimire for the careful reading and discussion of the manuscript. This work was supported by the National Institute of Health [J.K. R01-GM066225, R01-GM097074].

References

[1] M.S. Bartolomei, A.C. Ferguson-Smith, Mammalian genomic imprinting, Cold Spring Harb. Perspect. Biol. 3 (7) (2011).
[2] D.P. Barlow, M.S. Bartolomei, Genomic imprinting in mammals, Cold Spring Harb. Perspect. Biol. 6 (2) (2014).
[3] C.A. Edwards, A.C. Ferguson-Smith, Mechanisms regulating imprinted genes in clusters, Curr. Opin. Cell Biol. 19 (2007) 281–289.
[4] A.T. Hark, C.J. Schoenherr, D.J. Katz, R.S. Ingram, J.M. Levorse, S.M. Tilghman, CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus, Nature 405 (2000) 486–489.
[5] A.C. Bell, G. Felsenfeld, Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene, Nature 405 (2000) 482–485.
[6] J. Kim, A. Kollhoff, A. Bergmann, L. Stubbs, Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3, Hum. Mol. Genet. 12 (2003) 233–245.
[7] J.D. Kim, A.K. Hinz, A. Bergmann, J.M. Huang, I. Ovcharenko, L. Stubbs, J. Kim, Identification of clustered YY1 binding sites in imprinting control regions, Genome Res. 16 (2006) 901–911.
[8] J.D. Kim, S. Yu, J.H. Choo, J. Kim, Two evolutionarily conserved sequence elements for Peg3/Usp29 transcription, BMC Mol. Biol. 9 (2008) 108.
[9] H. He, A. Ye, B.P.U. Perera, J. Kim, YY1's role in the Peg3 imprinted domain, Sci. Rep. 7 (1) (2017) 6427.
[10] C.J. Schoenherr, J.M. Levorse, S.M. Tilghman, CTCF maintains differential methylation at the Igf2/H19 locus, Nat. Genet. 33 (2003) 66–69.
[11] N. Engel, J.L. Thorvaldsen, M.S. Bartolomei, CTCF binding sites promote transcription initiation and prevent DNA methylation on the maternal allele at the imprinted H19/Igf2 locus, Hum. Mol. Genet. 15 (2006) 2945–2954.
[12] J. Kim, Multiple YY1 and CTCF binding sites in imprinting control regions, Epigenetics 3 (2008) 115–118.
[13] A.P. Bird, CpG-rich islands and the function of DNA methylation, Nature 321 (6067) (1986) 209–213.
[14] A. Hodgkinson, A. Eyre-Walker, Variation in the mutation rate across mammalian genomes, Nat. Rev. Genet. 12 (11) (2011) 756–766.
[15] A.P. Wolffe, Chromatin: Structure and Function, 3rd ed., Academic Press, San Diego, California, 1998.